**Báo cáo**

# Đánh giá tập dữ liệu IRIS dựa trên ba thuật toán của cây quyết định

## J48

* Thuật toán đã sử dụng: J48
* Tên quan hệ: iris
* Số trường hợp trong mối quan hệ: 150
* Số thuộc tính: 5
* Danh sách thuộc tính: sepallength, sepalwidth, petallength, petalwidth, class
* The test mode you selected: split=66%
* Cây J48 đã được cắt tỉa
* Số lá: 5
* Kích thước cây: 9
* Thời gian build model: 0,01s
* Phân lớp chính xác: 96.0784 %
* Phân lớp không chính xác: 3.9216 %
* Ma trận nhầm lẫn:

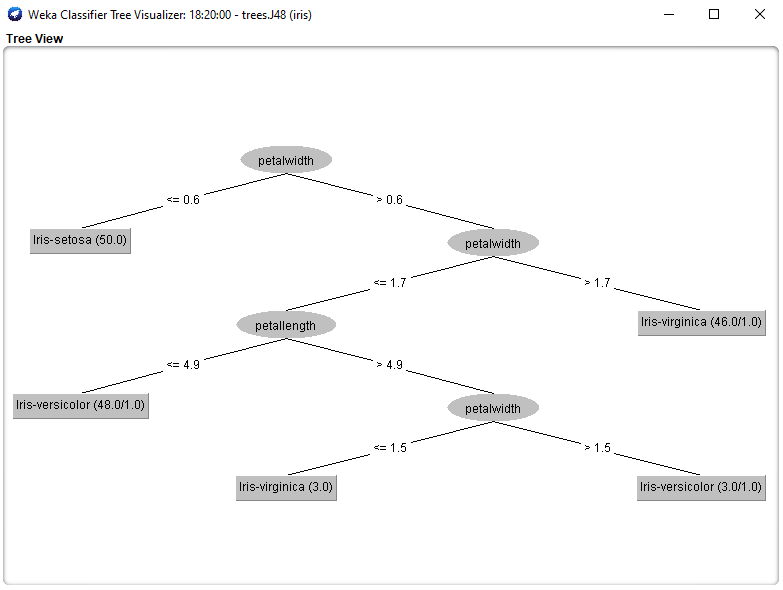
a b c <-- classified as

15 0 0 | a = Iris-setosa

0 19 0 | b = Iris-versicolor

0 2 15 | c = Iris-virginica

* Trực quan hoá cây



## RandomTree

* Thuật toán đã sử dụng: RandomForest
* Tên quan hệ: iris
* Số trường hợp trong mối quan hệ: 150
* Số thuộc tính: 5
* Kích thước cây: 17
* Danh sách thuộc tính: sepallength, sepalwidth, petallength, petalwidth, class
* The test mode you selected: split=66%
* Thời gian build model: 0s
* Phân lớp chính xác: 92.1569 %
* Phân lớp không chính xác: 7.8431 %
* Ma trận nhầm lẫn:

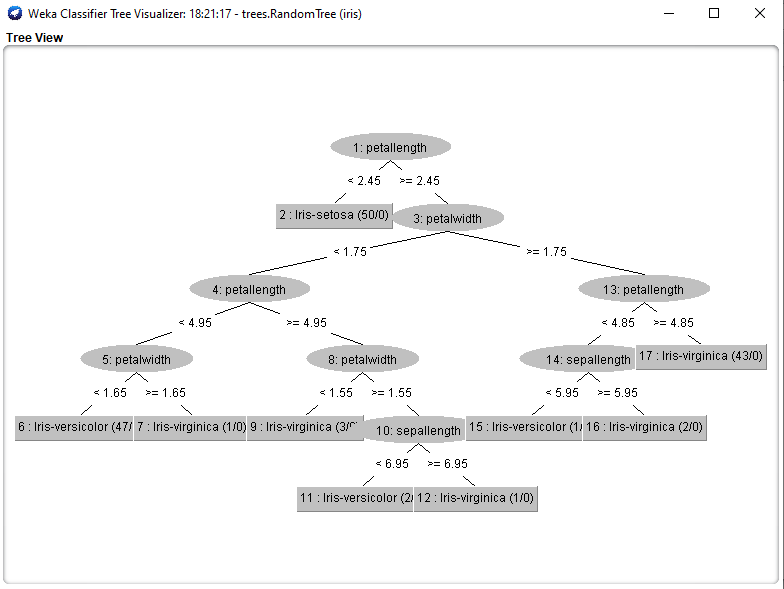
a b c <-- classified as

15 0 0 | a = Iris-setosa

0 17 2 | b = Iris-versicolor

0 2 15 | c = Iris-virginica

* Trực quan hoá cây



## REPTree

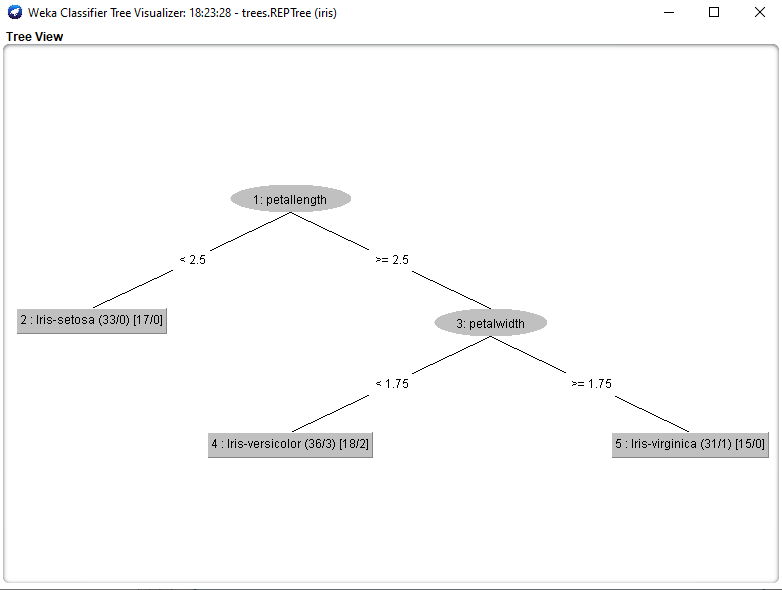
* Thuật toán đã sử dụng: REPTree
* Tên quan hệ: iris
* Số trường hợp trong mối quan hệ: 150
* Số thuộc tính: 5
* Kích thước cây: 5
* Danh sách thuộc tính: sepallength, sepalwidth, petallength, petalwidth, class
* The test mode you selected: split=66%
* Phân lớp chính xác: 92.1569 %
* Phân lớp không chính xác: 7.8431 %
* Ma trận nhầm lẫn:

a b c <-- classified as

15 0 0 | a = Iris-setosa

0 15 4 | b = Iris-versicolor

0 0 17 | c = Iris-virginica



# Đánh giá tập dữ liệu LABOR dựa trên DT, K-NN, SVM, NB

## Mô tả dữ liệu

* Title: Final settlements in labor negotitions in Canadian industry
* Thông tin nguồn: -- Creators: Collective Barganing Review, montly publication, Labour Canada, Industrial Relations Information Service, Ottawa, Ontario, K1A 0J2, Canada, (819) 997-3117. The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc) in Canada in 87 and first quarter of 88.
* -- Donor: Stan Matwin, Computer Science Dept, University of Ottawa, 34 Somerset East, K1N 9B4, (stan@uotcsi2.bitnet)
* -- Date: November 1988
* Số Instances: 57
* Số thuộc tính: 16
* Missing Attribute Values: None
* Bài toán phân lớp
* Thông tin thuộc tính

duration

wage-increase-first-year

wage-increase-second-year

wage-increase-third-year

cost-of-living-adjustment

working-hours

pension

standby-pay

shift-differential

education-allowance

statutory-holidays

vacation

longterm-disability-assistance

contribution-to-dental-plan

bereavement-assistance

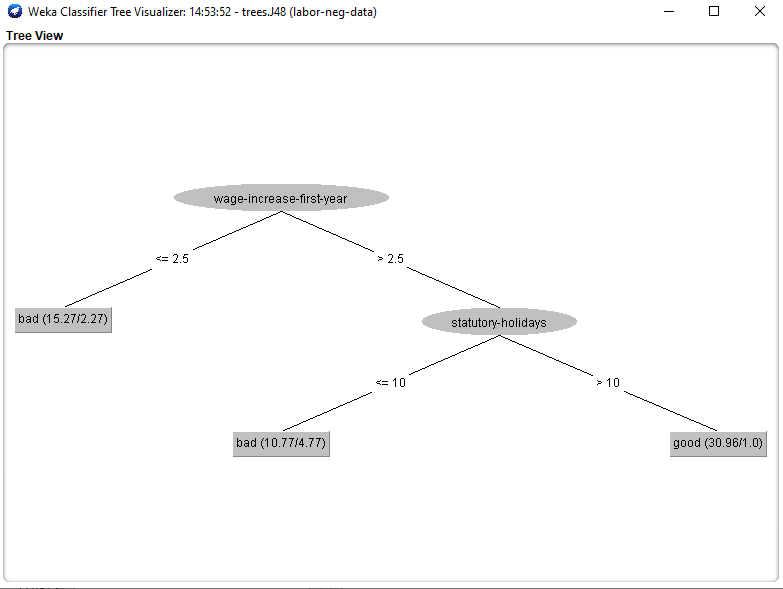
contribution-to-health-plan

class

* Mục tiêu bài toán: chọn ra thuật toán phân lớp tốt nhất, độ chính xác cao nhất trên tập dữ liệu labor có sẵn trong weka

## DT

| Thời gian build model | 0s | |
| --- | --- | --- |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 17 | |
| Số mẫu phân lớp sai | 2 | |
| Tỷ lệ phân lớp đúng | 89.4737 % | |
| Tỷ lệ phân lớp sai | 10.5263 % | |
| Số lá | 3 | |
| Kích thước cây | 5 | |
| Độ sâu cây | 2 | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  5 1 | a = bad  1 12 | b = good | |
|
|
|
|
|



## KNN

Với K = 1

| Thời gian build model | 0s | |
| --- | --- | --- |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 15 | |
| Số mẫu phân lớp sai | 4 | |
| Tỷ lệ phân lớp đúng | 78.9474 % | |
| Tỷ lệ phân lớp sai | 21.0526 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  6 0 | a = bad  4 9 | b = good | |
|
|
|
|
|

Với K = 3

| Thời gian build model | 0s | |
| --- | --- | --- |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 16 | |
| Số mẫu phân lớp sai | 3 | |
| Tỷ lệ phân lớp đúng | 84.2105 % | |
| Tỷ lệ phân lớp sai | 15.7895 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  6 0 | a = bad  3 10 | b = good | |
|
|
|
|
|

Với K = 9

| Thời gian build model | 0s | |
| --- | --- | --- |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 18 | |
| Số mẫu phân lớp sai | 1 | |
| Tỷ lệ phân lớp đúng | 94.7368 % | |
| Tỷ lệ phân lớp sai | 5.2632 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  5 1 | a = bad  0 13 | b = good | |
|
|
|
|
|

## SVM

| Thời gian build model | 0s | |
| --- | --- | --- |
| Sử dụng hàm nhân tuyến tính | | |
| Số Kernel | 1 | |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 17 | |
| Số mẫu phân lớp sai | 2 | |
| Tỷ lệ phân lớp đúng | 89.4737 % | |
| Tỷ lệ phân lớp sai | 10.5263 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  5 1 | a = bad  1 12 | b = good | |
|
|
|
|
|

## NB

| Thời gian build model | 0s | |
| --- | --- | --- |
| Tổng số bản ghi | 57 | |
| Tổng số lớp | 17 | |
| Số mẫu phân lớp đúng | 18 | |
| Số mẫu phân lớp sai | 1 | |
| Tỷ lệ phân lớp đúng | 94.7368 % | |
| Tỷ lệ phân lớp sai | 5.2632 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  5 1 | a = bad  0 13 | b = good | |
|
|
|
|
|

## Đánh giá

Bộ phân lớp tốt nhất là NaiveBayes và KNN k = 9 có độ chính xác cao nhất

# Phân cụm trên tập CUSTOMERS

## Đánh giá

* Relation: customers
* Instances: 9
* Thuộc tính: 5

income

age

children

education

* Test mode: Classes to clusters evaluation on training data
* cụm bị bỏ qua ‘marital\_status’ là một thuộc tính phân cụm được thực hiện trên.
* Số lần lặp lại: K = 3
* Trong cụm tổng các lỗi bình phương: 3,8499264974433896
* Điểm xuất phát ban đầu(random)

Cluster 0: 25000,35,3,high\_school

Cluster 1: 30000,20,0,high\_school

Cluster 2: 20000,40,0,high\_school

* Clustered Instances

0 4 ( 44%)

1 4 ( 44%)

2 1 ( 11%)

* Thuộc tính lớp: marital\_status
* Classes to Clusters:

0 1 2 <-- assigned to cluster

0 1 1 | single

2 2 0 | married

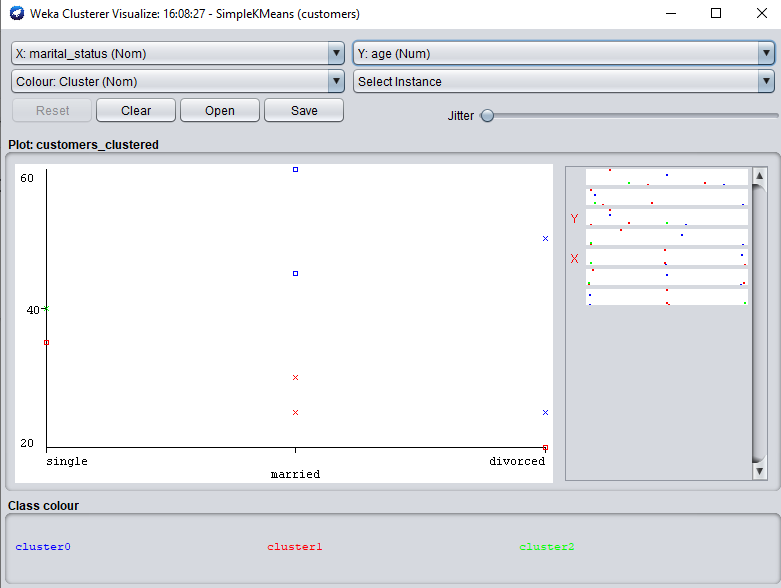
2 1 0 | divorced

Cluster 0 <-- divorced

Cluster 1 <-- married

Cluster 2 <-- single

* Các trường hợp được nhóm không chính xác: 4,0 44,4444%



# Phân cụm trên tập BANK

## Mô tả dữ liệu

Tên dữ liệu: bank

Nguồn dữ liệu: https://github.com/bluenex/WekaLearningDataset/blob/master/bank/bank.arff

| ***Missing*** | ***Mẫu*** | ***Thuộc tính*** | ***Kiểu dữ liệu*** |
| --- | --- | --- | --- |
| 0% | 300 | 9 | Nominal |

Danh sách thuộc tính: age, sex, region, income, married, children, car, mortgage, pep

## Mô tả bài toán

Bài toán phân cụm các khu vực

## Thuật toán

SimpleKmeans

## Đánh giá

* Số lần lặp lại: 4
* Trong tổng số lỗi bình phương trong cụm: 473.14244083827697
* Các trường hợp được nhóm không chính xác: 211,0 70,333%

=> Không khả dụng

* Thời gian thực hiện model: 0,01s
* Điểm xuất phát ban đầu (ngẫu nhiên):

Cluster 0: 26,FEMALE,16518.6,YES,NO,YES,NO,YES

Tuổi: 26

Giới tính: nữ

Thu nhập: 16518.6

Hôn nhân: có

Con cái: không

Ô tô: có

Thế chấp: không

Kế hoạch trả nợ: Có

Cluster 1: 35,FEMALE,33028.3,NO,YES,NO,NO,YES

Tuổi: 35

Giới tính: nữ

Thu nhập: 33028.3

Hôn nhân: không

Con cái: có

Ô tô: không

Thế chấp: không

Kế hoạch trả nợ: Có

Cluster 2: 22,MALE,7304.2,NO,NO,YES,YES,NO

Tuổi: 22

Giới tính: nam

Thu nhập: 7304.2

Hôn nhân: không

Con cái: không

Ô tô: có

Thế chấp: có

Kế hoạch trả nợ: không

Cluster 3: 37,MALE,24814.5,YES,YES,YES,YES,YES

Tuổi: 37

Giới tính: nam

Thu nhập: 24814.5

Hôn nhân: có

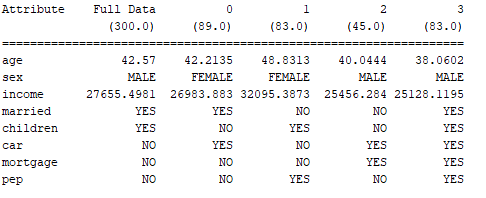
Con cái: có

Ô tô: có

Thế chấp: có

Kế hoạch trả nợ: có

* Các trung tâm cụm cuối cùng:



* Clustered Instances:

0 89 ( 30%)

1 83 ( 28%)

2 45 ( 15%)

3 83 ( 28%)

* Thuộc tính class: region
* Cụm lớp

0 1 2 3 <-- assigned to cluster

42 35 16 44 | INNER\_CITY

16 16 8 11 | RURAL

23 23 20 21 | TOWN

8 9 1 7 | SUBURBAN

Cluster 0 <-- RURAL

Cluster 1 <-- SUBURBAN

Cluster 2 <-- TOWN

Cluster 3 <-- INNER\_CITY

